UNIVERSIDAD NACIONAL DE COLOMBIA

Machine Learning Lab 2.1.

Miguel Andrés Villamil Carrillo

Universidad Nacional de Colombia
Bogotá, Colombia
2018

1 Multinomial Naive Bayes Implementation.

The data set used is displayed in the next table.

Table 1: Table 1 (13.10)

	docID	words in document	in c = china?
Training set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
Test set	5	Taiwan Taiwan Sapporo	?

For the Multinomial Naive Bayes classifier implementation, the equation (1) will be applied to each of the classes (c) where c stands for "in China" and ($\bar{c}$) for "not in China"), using the Laplace smoothing technique.

$$P(c|d) \; \alpha \; P(c) \prod_{k=1}^{n} P(t_k|c) \qquad (1)$$

The equation (2) is used to calculate each of the ($P(t_k|c)$) on the data set.

$$P(t|c) = \frac{T_{ct}+1}{\sum_{t\epsilon V} T_{ct}+B} \qquad (2)$$

With ($B = 7$) in the equation (2) and ($P(c) = P(\bar{c}) = \frac{1}{2}$) in the equation (1). Applying the equation (2) to the training set we end with the next results.

$$\begin{array}{lll} P(Taiwan|c) & = (2+1)/(5+7) & = 1/4 \\ P(Sapporo|c) & = ((0+1)/(5+7)) & = (1/12) \\ P(Taiwan|\bar{c})) & = (1+1)/(5+7) & = 1/6 \\ P(Sapporo|\bar{c}) & = (2+1)/(5+7) & = 1/4 \end{array}$$

Based on the previous results when attempting to classify the test set we get the results:

$$\begin{array}{lll} P(c|d_5) & \alpha \; \frac{1}{2} * (\frac{1}{4})^2 * \frac{1}{12} & \approx 0.00260 \\ P(\bar{c}|d_5) & \alpha \; \frac{1}{2} * (\frac{1}{6})^2 * \frac{1}{4} & \approx 0.00347 \end{array}$$

By analizing the results we can say that the test set is more likely to not be in c.

2 Bernoulli NB Implementation.

Based on Table 2 the values for $P(c|d_5)$ and $P(\bar{c}|d_5)$ are:

$P(c|d_5) = \frac{1}{2} * \frac{3}{4} * \frac{1}{4} * (1 - \frac{1}{2}) * (1 - \frac{1}{2}) * (1 - \frac{1}{2}) * (1 - \frac{1}{4}) * (1 - \frac{1}{4})$

$P(c|d_5) \approx$0.00659

Table 2:

Term	c	$\bar{c})$
Taipei	$(1+1)/(2+2) = 1/2$	$(0+1)/(2+2) = 1/4$
Taiwan	$(2+1)/(2+2) = 3/4$	$(1+1)/(2+2) = 1/2$
Macao	$(1+1)/(2+2) = 1/2$	$(0+1)/(2+2) = 1/4$
Shanghai	$(1+1)/(2+2) = 1/2$	$(0+1)/(2+2) = 1/4$
Japan	$(0+1)/(2+2) = 1/4$	$(1+1)/(2+2) = 1/2$
Sapporo	$(0+1)/(2+2) = 1/4$	$(1+1)/(2+2) = 1/2$
Osaka	$(0+1)/(2+2) = 1/4$	$(1+1)/(2+2) = 1/2$

P(—d5)= 1$\frac{}{2*\frac{1}{2}*\frac{1}{2}*(1-\frac{1}{4})*(1-\frac{1}{4})*(1-\frac{1}{4})*(1-\frac{1}{2})*(1-\frac{1}{2})}$

$P(\bar{c}|d_5) \approx 0.01318$

The test set is more likely to not be in (c).